NCTCIR-9 GeoTime: Geo-temporal Information Retrieval Based on Semantic Role Labeling and Rank Aggregation

Yoonjae Jeong, Gwan Jang, Kyung-min Kim and Sung-Hyon Myaeng
Korea Advanced Institute of Science and Technology (KAIST)
NTCIR-9 GeoTime Task
- It is about geographic and temporal search in news articles.
- We participated in the English sub-task only.

Topics asks for geographic- and temporal-based information.
- A topic asks for information on where and when a particular event occurred or what event happened at a specific time and location.
- “where, when, and what did <entities> <action>?"”

The proposed geo-temporal information retrieval model
- Our basic idea is to add locational and temporal aspects to terms in a document using Semantic Role Labeling (SRL).
Semantic Roles

- A semantic role is the underlying relationship that a participant (linguistic constituent) has with the main verb in a clause.
- In the form, the elements of topic are reflected in following semantic roles.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Semantic Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>where</td>
<td>AM-LOC</td>
</tr>
<tr>
<td>when</td>
<td>AM-TMP</td>
</tr>
<tr>
<td>&lt;entity&gt;</td>
<td>A0-5 (e.g., Agent, Patient, …)</td>
</tr>
<tr>
<td>&lt;action&gt;</td>
<td>verb</td>
</tr>
</tbody>
</table>
Overview of the proposed geo-temporal information retrieval

The Proposed Method

Overview of the proposed geo-temporal information retrieval

Document Representations

Corpus

Preprocessing Documents

Preprocessing Topics

Topic Representations

Rank Aggregation

Basic Document Language Model (BDLM)
Role-based Document Language Model (RDLM)
Basic Sentence Language Model (BSLM)
Role-based Sentence Language Model (RSLM)

Topic-Document Scores
Documents are represented as sets of words for semantic roles.

- An example of SRL for a document

[A-1 Astrid Lindgren, the Swedish writer whose rollicking, anarchic books about Pippi Longstocking horrified a generation of parents and captivated millions of children around the globe], died in her sleep [AM-TMP Monday] [AM-LOC at her home in Stockholm, Sweden.]

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_V$</td>
<td>A set of verb in document</td>
</tr>
<tr>
<td></td>
<td>e.g., die</td>
</tr>
<tr>
<td>$T_A$</td>
<td>A set of terms with numbered argument roles (A0-5) in document</td>
</tr>
<tr>
<td></td>
<td>e.g., Astrid, Lindgren, ... , children, globe</td>
</tr>
<tr>
<td>$T_{AM-LOC}$</td>
<td>A set of terms with location (AM-LOC) roles in document</td>
</tr>
<tr>
<td></td>
<td>e.g., home, Stockholm, Sweden</td>
</tr>
<tr>
<td>$T_{AM-TMP}$</td>
<td>A set of terms with temporal role (AM-TMP) in document</td>
</tr>
<tr>
<td></td>
<td>e.g., Monday</td>
</tr>
</tbody>
</table>
The Proposed Method
Topic Representations (1/2)

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Q-LOC</strong></td>
<td>Whether a question is about location or not? e.g., <em>When and where did Astrid Lindgren die?</em></td>
</tr>
<tr>
<td><strong>Q-TMP</strong></td>
<td>Whether a question is about time or not? e.g., <em>When and where did Hurricane Katrina make landfall in the United States?</em></td>
</tr>
<tr>
<td><strong>Q-AGT</strong></td>
<td>Whether is a question about agent or not? e.g., <em>What Portuguese colony was transferred to China and when?</em></td>
</tr>
<tr>
<td><strong>Q-MSC</strong></td>
<td>The others e.g., <em>How old was Max Schmeling when he died, and where did he die?</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set of vocabularies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_v$</td>
<td>A set of vocabularies in verb role in topic</td>
</tr>
<tr>
<td>$V_a$</td>
<td>A set of vocabularies in numbered argument (A0-5) roles in topic</td>
</tr>
<tr>
<td>$V_{AM-LOC}$</td>
<td>A set of vocabularies in locational role (AM-LOC) in topic</td>
</tr>
<tr>
<td>$V_{AM-TMP}$</td>
<td>A set of vocabularies in temporal role (AM-TMP) in topic</td>
</tr>
</tbody>
</table>
To determine the question types, we devised some heuristic rules based on syntactic parser results.

- A parse-tree example for question type identification

**Topic:** When and where did Astrid Lindgren die?

**Parsing Tree:**

```
(ROOT
  (SBARQ
    (WHADVP (WRB When))
    (CC and)
    (WRB where))
  (SQ (VBD did))
  (NP (NNP Astrid) (NNP Lindgren))
  (VP (VB die)))
(.) ?)))
```
Basic Document Language Model (BDLM)

- Basic Language Model for a topic $q$ and document $d$.

$$P_{BDLM}(d|q) = P(q|d) \times \frac{P(d)}{P(q)} \approx \prod_{t \in q} P(t|d)$$

$$P(t|d) = \frac{tf_{t,d} + \mu \times P(t|D)}{|d| + \mu}$$

- $d$: a document
- $q$: a topic
- $t$: a term in $q$
- $tf_{t,d}$: the term frequency of term $t$ in document $d$
- $D$: the set of all document in the corpus
- $\mu$: the smoothing parameter ($=2500$)
Role-based Document Language Model (RDLM)

- Based on the document representation, we built the RDLM where $R$ represents the semantic roles in the document $d$.

$$P_{RDL}m (d|q) = \prod_{r \in R} \left( P(q_r|d_r) + \alpha \right)$$

\[
\alpha = \begin{cases} 
1, & \text{if Q-LOC is true, } r=\text{AM-LOC and } |d_r| > 0. \\
1, & \text{if Q-TMP is true, } r=\text{AM-TMP and } |d_r| > 0. \\
0, & \text{otherwise}
\end{cases}
\]  

(3)

$R$: the semantic roles in $q$

$q_r$: the set of terms given the role $r$ in $q$

$d_r$: the set of terms given the role $r$ in $d$
Basic Sentence Language Model (BSLM)

- Sometimes the relevant information related to a topic is fully contained in one sentence in document.

\[
P_{\text{BSLM}}(d|q) = \max_{s \in S} P(s|q) = \max_{s \in S} \prod_{t \in q} P(t|s) \tag{4}
\]

- \(S\) : a set of sentence in \(d\)
- \(s\) : a sentence in \(S\)
Role-based Sentence Language Model (RSLM)

- RSLM adds semantic roles to BSLM in the same way RDLM was constructed out of BDLM.

\[
P_{RSLM}(d|q) = \max_{s \in S} \prod_{r \in R} \left( P(q_r|s_r) + \alpha \right)
\]

\[
\alpha = 1, \text{ if Q-LOC is true, } r=\text{AM-LOC and } |s_r| > 0.
\]

\[
\alpha = 1, \text{ if Q-TMP is true, } r=\text{AM-TMP and } |s_r| > 0.
\]

\[
\alpha = 0, \text{ otherwise}
\]

\( R \): the semantic roles in \( q \)

\( q_r \): the set of terms given the role \( r \) in \( q \)

\( s_r \): the set of terms given the role \( r \) in \( s \)
### Relevant documents and their ranks and normalized scores of each model for Topic GeoTime-0025 in NTCIR-8 corpus

<table>
<thead>
<tr>
<th>Relevant Document</th>
<th>Rank &amp; Normalized Score</th>
<th>Rank &amp; Normalized Score</th>
<th>Rank &amp; Normalized Score</th>
<th>Rank &amp; Normalized Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NYT_ENG_20041226.0096</strong></td>
<td>24 (8.95E-01)</td>
<td>201 (1.54E+00)</td>
<td>251 (3.72E-02)</td>
<td>4 (2.82E+02)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20041229.0208</strong></td>
<td>167 (1.07E-04)</td>
<td>322 (1.07E+00)</td>
<td>133 (9.18E-02)</td>
<td>3 (2.86E+02)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20041230.0186</strong></td>
<td>3 (2.20E+02)</td>
<td>18 (4.46E+01)</td>
<td>298 (3.09E-03)</td>
<td>102 (7.74E-01)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20041230.0204</strong></td>
<td>8 (8.47E+01)</td>
<td>26 (3.26E+01)</td>
<td>302 (3.09E-03)</td>
<td>98 (7.75E-01)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20041230.0245</strong></td>
<td>4 (1.21E+02)</td>
<td>17 (4.46E+01)</td>
<td>299 (3.09E-03)</td>
<td>101 (7.74E-01)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20041230.0256</strong></td>
<td>6 (1.20E+02)</td>
<td>25 (3.49E+01)</td>
<td>303 (3.09E-03)</td>
<td>100 (7.74E-01)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20041231.0009</strong></td>
<td>2 (2.20E+02)</td>
<td>19 (4.46E+01)</td>
<td>300 (3.09E-03)</td>
<td>99 (7.75E-01)</td>
</tr>
<tr>
<td><strong>NYT_ENG_20050328.0205</strong></td>
<td>36 (4.30E-03)</td>
<td>349 (1.07E+00)</td>
<td>88 (9.18E-02)</td>
<td>162 (7.72E-01)</td>
</tr>
</tbody>
</table>
The proposed IR models have difference characteristics.

It can be combined to handle various retrieval cases by devising a rank aggregation method, which is based on Dwork, et al. (2001)’s Markov Chain based approaches (MC1, MC2, MC3, & MC4).

We adopted MC2 because it is arguably the most representative of minority viewpoints of sufficient statistical significance; it protects specialist views.
The transition matrix of the $k^{th}$ rank list $T_k$, $T_k$ is

$$T_k \triangleq \left(t_{ij}^{(k)}\right)_{mn}$$

(6)

$$i_{ij}^{(k)} = \begin{cases} 
\frac{1}{\# \{j \mid j > \tau_k i \text{ or } i = i \}}, & j > \tau_k i \text{ or } j = i \\
0, & \text{otherwise}
\end{cases}$$

(7)

$j > \tau_k i$: document $j$ is ranked higher than document $i$ in ranking list $T_k$.

The final transition matrix $T$ is

$$T = \frac{1}{l} \sum_{k=1}^{l} T_k$$

(8)

$l$: the number of ranked lists
The effective ranks for aggregations of RDLM, BSLM, and RSLM are a small number of top ones.

We applied the threshold $\theta$ to the elements of transition matrix.

$$t^{(k)}_{ij} = \begin{cases} 
1 & \#\{j \mid j > \tau_k \text{ or } i = i\} > \theta \\
0, & \text{otherwise}
\end{cases}$$

The final score is

$$x = T^T x_0, \quad x_0 = \begin{bmatrix}
1/|D| \\
\vdots \\
1/|D|
\end{bmatrix}$$
Evaluation
Test in NTCIR-8 Corpus

BDLM and the Rank Aggregations with threshold $\theta$

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>nDCG@100</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
<td>0.4087</td>
</tr>
<tr>
<td>$\theta$=0</td>
<td>0.4688</td>
</tr>
<tr>
<td>$\theta$=50</td>
<td>0.4684</td>
</tr>
<tr>
<td>$\theta$=100</td>
<td>0.4885</td>
</tr>
<tr>
<td>$\theta$=150</td>
<td>0.4883</td>
</tr>
<tr>
<td>$\theta$=200</td>
<td>0.4911</td>
</tr>
<tr>
<td>$\theta$=250</td>
<td>0.4925</td>
</tr>
<tr>
<td>$\theta$=300</td>
<td>0.4911</td>
</tr>
</tbody>
</table>
### Summited Runs

<table>
<thead>
<tr>
<th>RUN</th>
<th>Topic Source</th>
<th>Aggregation</th>
<th>Aggregation Threshold (θ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRNLP-EN-EN-1-D</td>
<td>Description only</td>
<td>(BD, RD, BS, &amp; RS) LM</td>
<td>150</td>
</tr>
<tr>
<td>IRNLP-EN-EN-2-D</td>
<td>Description only</td>
<td>(BD, RD, BS, &amp; RS) LM</td>
<td>200</td>
</tr>
<tr>
<td>IRNLP-EN-EN-3-DN</td>
<td>Description &amp; Narrative</td>
<td>(BD, RD, BS, &amp; RS) LM</td>
<td>200</td>
</tr>
<tr>
<td>IRNLP-EN-EN-4-DN</td>
<td>Description &amp; Narrative</td>
<td>(BD, RD, BS, &amp; RS) LM</td>
<td>140</td>
</tr>
<tr>
<td>BDLM-D</td>
<td>Description only</td>
<td>BDLM</td>
<td>-</td>
</tr>
<tr>
<td>BDLM-DN</td>
<td>Description &amp; Narrative</td>
<td>BDLM</td>
<td>-</td>
</tr>
</tbody>
</table>
Evaluation
Results from NTCIR-9 (1/2)

MAP  nDCG @ 10  nDCG @ 100  nDCG @ 1000

- BDLM-D
- IRNLP-EN-EN-01-D
- IRNLP-EN-EN-02-D
- BDLM-DN
- IRNLP-EN-EN-03-DN
- IRNLP-EN-EN-04-DN
Topics showing high performances

- Those verbs are related to the activities or states of agents clearly (e.g. “murder”, “hijack”, “kill”, and so on).
- The terms are also not ambiguous because they are proper nouns or very specific number of theme (e.g. “4 people” in GeoTime-0033).

Topics showing low performances

- Many errors in the analysis of topics
- The verbs were related to the existence or occurrence of agent or theme. (e.g., “occur”, “happen”)
- They sometimes require inference or term expansion.
A new geo-temporal information retrieval method that utilizes **semantic role labeling** and **rank aggregation**.

- It is useful to analyze documents for semantic roles around main predicates of sentences and generate language models after the analysis.
- While the SRL-based method is not always superior across different topics, they complement the usual language modeling approach.
- It warrant the proposed rank aggregation method.

**Future Work**

- Term expansion and weighting are necessary.
- It is also needed to find the optimal weight and thresholds in rank aggregation.
Thank you for your attention.